Unit 5 – Regression and Correlation
Practice Problems (2 of 3)

Due: Thursday October 24, 2024
Last date to submit late for credit (-20 points): Thursday October 31, 2024

Before you begin. Download from the course website
hersdata_small.xlsx

Description of Dataset

Source
Hulley et al (1998) Randomized trial of estrogen plus progestin for secondary prevention of heart disease in postmenopausal women. The Heart and Estrogen/progestin Replacement Study. *Journal of the American Medical Association*, **280**(7), 605-613

The Heart and Estrogen/progestin Replacement Study (HERS) was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus.

The dataset for this homework (**hersdata_small.xlsx**) is a simple random sample of n=1000. A subset of the variables are considered:

Data dictionary/Codebook *(Partial)*

Variable	Label	Type	Codings
age	Age, years	numeric	Continuous, range, [45:79]
BMI	Body Mass index (kg/m²)	numeric	Continuous, range, [15.21:54.13]
glucose	Fasting glucose (mg/dL)	numeric	Continuous, range, [29:298]
LDL	LDL cholesterol (mg/dL)	numeric	Continuous, range, [44.4:393.4]
drinkany	Any current alcohol use	numeric	1 = yes 0 = no
exercise	Exercise at least 3x/week	numeric	1 = yes 0 = no
HT	Randomization	numeric	1 = hormone therapy 0 = placebo
physact	Comparative ("compared to other women your age") physical activity	Numeric	1 = much less active 2 = somewhat less active 3 = about as active 4 = somewhat more active 5 = much more active
statins	Statin use	Numeric	1 = yes 0 = no
diabetes	Diabetes	Numeric	1 = yes 0 = no

 # 1.
This exercise gives you practice with doing a preliminary step in regression: inspecting your data.

By any means you like, obtain numerical summaries of the four continuous variables: **`age, BMI, glucose,`** and **`LDL.`**

 # 2.
This exercise also gives you practice with doing a preliminary step in regression: inspecting your data.

By any means you like, obtain numerical summaries of the six discrete variables: **`drinkany, exercise, HT, physact, statins, diabetes.`**

Exercises #3 - #7 consider non-diabetics only (diabetes==0)

#3
This exercise gives you practice with creating a subset of your data and fitting and intepreting a single predictor model.

Fit a single predictor model of Y=**`glucose`** to X= **`exercise`** among non-diabetics ONLY. In 1-2 sentences, report and interpret the output.

#4
This exercise gives you practice with fitting and intepreting a multiple predictor model.

Next fit a multiple predictor model of Y= **`glucose`** among non-diabetics ONLY.. Fit the following predictors: **`exercise, age, drinkany,`** and **`BMI`**. In 1-2 sentences, interpret the output.

#5
This exercise gives you practice comparing two hierarchical (nested) models with a partial F test.

Perform a partial F-test for the significance of **`exercise`** controlling for **`age,drinkany,`** and **`BMI`** among non-diabetics ONLY. Interpret.

#6

This exercise gives you practice creating design variables that are a set of 0/1 indicator variables.

Create four 0/1 design variables to represent the 5 possible outcomes of **`physact`** among non-diabetics ONLY. By any means you like, produce a check on the creation of your design variables.

#7

This exercise gives you practice including design variables in a regression model and then interpreting their fit.

Fit a multiple predictor model of Y= **`glucose`** among non-diabetics ONLY. Consider as the predictor ONLY the design variables for **`physact`**. In 1-2 sentences, interpret the output.